DOCUMENT RESUME

ED 159 220 TM 007 702

AUTHOR Coffman, William E.; Shigemasu, Kazuo
TITLE Appraising School Effectiveness Using a Bayesian Method.
PUB DATE 29 Mar 78
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Gains; *Analysis of Covariance; *Bayesian Statistics; Curriculum Evaluation; Educational Accountability; *Educational Assessment; *Elementary Schools; Intermediate Grades; *Mathematical Models; *Measurement Techniques; Program Evaluation; Sampling; Standard Error of Measurement; Tests of Significance
IDENTIFIERS Sample Size; *School Evaluation

ABSTRACT

Appraisal of a school's relative effectiveness is complicated by: (1) the need to control for input differences; (2) measurement error in input measures; and (3) small sample size within schools. This study compares the performance of two successive cohorts in 19 schools in a small midwestern city on the five Iowa Tests of Basic Skills using both a classical analysis of covariance and a Bayesian method. The Bayesian estimates were less different from cohort to cohort than the classical estimates and they were more consistent with expectations based on knowledge of school practices. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED159220

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

William E. Coffman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

APPRAISING SCHOOL EFFECTIVENESS USING A BAYESIAN METHOD

William E. Coffman
University of Iowa

and

Kazuo Shigemasu
Tohoku University

Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, March 29, 1978

PRINTED IN U.S.A.

TM007202

Appraising School Effectiveness Using a Bayesian Method

WILLIAM E. COFFMAN, University of Iowa and

KAZUO SHIGEMASU, The National University of Tohoku

ABSTRACT

Appraisal of a school's relative effectiveness is complicated by (1) the need to control for input differences, (2) measurement error in input measures, and (3) small sample size within schools. This study compares the performance of two successive cohorts in 19 schools in a small midwestern city on the five Iowa Tests of Basic Skills using both a classical analysis of covariance and a Bayesian method. The Bayesian estimates were less different from cohort to cohort than the classical estimates and they were more consistent with expectations based on knowledge of school practices. Suggestions for further study are given.

PRINTED IN U.S.A.

Appraising School Effectiveness Using a Bayesian Method

WILLIAM E. COFFMAN, University of Iowa and
KAZUO SHIGEMASU, National University of Tohoku

William E. Coffman
334 Lindquist Center
University of Iowa
Iowa City, IA 52242
319-353-3354

More and more in recent years, educators are being called upon to render an account to the public from time to time indicating how well the school is fulfilling its task of guiding learning. Legislators are calling for accountability as a prerequisite to new appropriations. Newspapers seek information about the results of testing programs and publish the findings for all to read and interpret. To the layman's way of thinking, the task of rendering an account is a simple one: administer a commonly used battery of achievement tests and report the average scaled score or percentile rank for the school. Schools with high scores are obviously doing a good job; those with low scores must be doing something wrong.

To the professional educator, however, it is obvious that the direct comparison of such scores is misleading. The characteristics of pupils in one school may differ markedly from those in another school, and if the differences at the time a period of instruction begins are considered, it well may be that a school reporting lower average test scores at the end of a period of instruction may actually be achieving more impressive results than another school with higher terminal averages. In making a reasonable comparison among the measures for a group of schools, it is necessary to take account of differences in the characteristics of pupils in the several schools prior to instruction.

The question of how such adjustment might be made, however, is not easy to answer satisfactorily. Because of errors of measurement and of differences in expected gains depending on initial status, simple gain scores are usually unsatisfactory, particularly since the number of cases available in the typical school for establishing the averages is usually small. Furthermore, given the typical spread of ability within the typical school group, any average seldom tells the whole story. One would like to know what is happening to pupils at different levels of initial status.

This study is an early one in a planned series designed to provide a school effectiveness index for each school in order to stimulate study of teaching practices. It investigates the appropriateness of the analysis of covariance model for estimating adjusted mean scores for groups of schools not too widely separated on initial status, for whom it is reasonable to assume that regression slopes are not greatly different and where measures of the covariates are highly reliable. It compares two methods of estimating adjusted mean scores, the classical analysis of variance method and a Bayesian method that assumes exchangeability (DeFinetti, 1964). Approximately, the Bayesian model estimate of the adjusted mean α_1 is,

$$\tilde{\alpha}_i \simeq \tilde{\rho}_i \hat{\alpha}_i + (1 - \tilde{\rho}_i)\hat{\alpha}. \quad ,$$

where

$$\tilde{\rho}_i = \frac{\tilde{\phi}_\alpha}{\tilde{\phi}_\alpha + n_i^{-1}\tilde{\phi}} \quad ,$$

(~ over the parameters denotes the Bayesian model estimate and n_i is the number of pupils in the i-th school,)

$\hat{\alpha}_i$ is a classical estimate, and

$$\hat{\alpha}. = \frac{\Sigma n_i \hat{\alpha}_i}{\Sigma n_i} \quad .$$

The data for this analysis consist of test scores for two successive cohorts of pupils in 19 elementary schools in a midwestern city. The scores are for the five tests in Vocabulary (V), Reading (R), Language (L), Work Study Skills (W), and Mathematics (M) and for the Composite (C) for the Iowa Tests of Basic Skills administered in Grade 4 and Grade 6.

The analysis compares the two estimation methods and two different covariates, the five different tests taken in Grade 4 and the Grade 4 composite. Since there are two cohorts and two different sets of covariates as well as two estimation methods, there are eight different estimates for each posttest.

The Bayesian estimates are less different from cohort to cohort than the classical estimates. Furthermore, the Bayesian estimates correspond more nearly to our intuitions based on our knowledge of school practices. According to the classical estimates, the ranks of the six mean square of differences are M-L-V-R-W-C (with five subtests as covariates) and M-V-L-R-W-C with the composite score as covariate, whereas for the Bayesian estimates the ranks are L-M-W-C-V-R in both cases. Our feeling is that the subtests which are most resistant to change are Vocabulary, Reading, and Work Study: Vocabulary, because the pupils build vocabulary from a variety of sources other than school; Reading, because the reading programs are receiving primary emphasis in all the schools; and Work Study Skills, because no schools have specific courses that focus on these skills. Finally, if differences as large as 1.25 months from the overall average are considered to be of practical significance, the Bayesian estimates provide only one instance in which a school exhibits a positive deviation for one cohort and a negative deviation for the next cohort whereas for the classical estimates there are five instances, four of which are for vocabulary, an achievement judged to be particularly resistant to change as a result of instruciton.

Reports such as this should be viewed as suggestive only. Certainly three possible improvements, which are set forth in the complete paper, should be examined carefully before implementing the reporting system.

Appraising School Effectiveness Using a Bayesian Method

More and more in recent years, educators are being called upon to render an account to the public from time to time indicating how well the school is fulfilling its task of guiding the learning of a community's children. Legislators are calling for accountability as a prerequisite to new appropriations. Newspapers seek information about the results of testing programs and publish the findings for all to read and interpret. To the layman's way of thinking, the task of rendering an account is a simple one: administer a commonly used battery of achievement tests and report the average scaled scores or percentile ranks for each school. Schools with high values are obviously doing a good job; those with low scores must be doing something wrong.

To the professional educator, however, it is obvious that the direct comparison of the scores is misleading. The characteristics of pupils in one school may differ markedly from those in another school, and if the differences at the time a period of instruction begins are considered, it well may be that a school reporting lower average test scores at the end of a period of instruction may actually be achieving more impressive results than another school with higher terminal averages. In making a reasonable comparison among the measures for a group of schools, it is necessary to take account of differences in the characteristics of pupils in the several schools prior to instruction.

The question of how such adjustments might be made has been explored by many different investigators (e.g. Campbell & Stanley, 1963; Campbell & Erlebacher, 1970; Cronbach, et al., 1976; Cronbach & Furby, 1970; Dyer,

1970; Dyer, Linn & Patton, 1967; Horst, Tallmage & Wood, 1975; Lord, 1967; Porter, 1967), and the issue is far from settled. Essentially, all methods involve adjusting the final achievement measures using measures taken before the beginning of a period of instruction. Studies by Dyer, Linn & Patton (1969) and by Forsyth (1973) indicate that selective dropout may lead to invalid conclusions regarding school comparisons unless estimates are based on matched cases rather than on means computed from all cases present both at pretest and posttest administrations. This means that often results are based on relatively small samples. The validity of conclusions, therefore, will often depend to a considerable extent on the efficiency of the procedures used for estimating the adjusted output scores.

In this paper we will define school effectiveness in terms of the changes in pupil's scores on a standardized achievement test battery after some fixed interval. Of course we realize that changes in achievement test scores represent only one aspect of the outcome resulting from school activities; however, it is assumed that the battery has been selected as appropriate in relation to certain of the goals of the school and that the comparisons growing out of the analysis can provide a school staff and the administrators of a system with insights into the effects of the educational program so far as these goals are concerned. If the gains the pupils attain are relatively high, surely the school staff responsible for guiding pupil learning has been relatively successful. If, on the other hand, the gains are relatively low, the school staff might well reconsider the instructional procedures that have been used.

A direct comparison of pupil gains, while somewhat less misleading

than a direct comparison of the achievement test scores themselves, is still far from a satisfactory basis for comparing schools because it is not at all clear that the expected gains are equal for pupils with differing initial scores. The procedure for comparing the gains should make due allowance for preexisting differences among the pupil groups in the several schools. It should also minimize the effects of random fluctuations in the observations. The purpose of this paper is to report the results of our effort to produce a reliable index of school effectiveness.

Analysis of Covariance Model

Because the idea of comparing school effectiveness in terms of gain during the teaching process is a very natural one, we first consider the meaning of the comparison of gains. The problem of "measuring change" has been a controversial issue, and it will be necessary to make our position clear with respect to this issue.

Let us denote Y_{ij} as a post-test score of the j-th person in the i-th school and X_{ij1}, X_{ij2}, ..., X_{ijp} as the p covariates which are observed before the teaching interval of interest. Especially, X_{ij1} refers to a pre-test score of the j-th person in the i-th school, i.e. a parallel measure of the same function as measured by Y_{ij}. Then, the mean gain in the i-th school is,

$$E_j(Y_{ij} - X_{ij1}) \qquad (1)$$

where E_j means the expectation over persons j within group i. Sometimes, however, the interest is in estimating "true" changes of individuals.

In this case, i denotes an individual and j denotes a parallel test. Cronbach and Furby (1970) argued that gain scores are rarely useful and we agree that the estimation of (1) is not appropriate even for the specific purpose discussed in this paper. Usually there are considerable preexisting differences between schools in terms of the covariates, because, of course, pupils are not assigned randomly to schools. Rather than to make comparisons of achievement measures directly, therefore, it will be "fairer" to compare the expected gains, assuming that the location on the covariates are the same across the schools, say $X_{ij1} = C_1$, $X_{ij2} = C_2$ ...., $X_{ipj} = C_p$. That is,

$$E_j(Y_{ij} - X_{ij1} | X_{ij1} = C_1, \ldots, X_{ijp} = C_p). \qquad (2)$$

Because the value of X_{ij1} is fixed to be a common value for all groups, the comparison based on (2) is the same as that based on the following expected value:

$$E_j(Y_{ij} | X_{ij1} = C_1, \ldots, X_{ijp} = C_p). \qquad (3)$$

Thus, we are in a position to specify the form of the expected value of Y_{ij} with X_{ij1}, ...X_{ijp} given. Let us assume, for example, the analysis of covariance model,

$$E_j(Y_{ij} | X_{ij1}, \ldots, X_{ijp}) = \alpha_i + \beta_1 X_{ij1} + \ldots + \beta_p X_{ijp}, \qquad (4)$$

is appropriate. Then, because of the assumed homogeneity of regression coefficients across schools, it does not matter what values C_1, C_2, ..., C_p are, and the comparison over the whole range of the covariates reduces

to the comparison of the α_i. We will use α_i as an index of relative school effectiveness, and the purpose of this paper is to obtain a reliable estimate of α_i.

We assume that the analysis of covariance model is a reasonable approximation to the particular application treated in this paper. As is often pointed out, however, the analysis of covariance model is a subtle and delicate tool, and it requires an investigator's careful consideration whether or not this is a valid model (Cronbach et al., 1976; Lindquist, 1953; Elashoff, 1969). Four assumptions are involved in the analysis of covariance model, i.e.,

(i) The relationships between Y_{ij} and the covariates are linear,

(ii) The regression coefficients are homogeneous among schools,

(iii) With X_{ij1}, ..., X_{ijp} given, Y_{ij} is independently distributed according to a normal distribution,

(iv) The variance of the distribution of Y_{ij} conditional on X_{ij1}, ..., X_{ijp} is the same for all schools and all possible X values.

Experiences with educational data in situations similar to that in this paper indicate that it is often reasonable to assume (i) and (iii). But it is more difficult to tell a priori whether or not the residual variances or the regression coefficients are homogeneous. Our feeling is that some statistical method of checking homogeneity should be involved in the process of analysis of covariance procedure. Usually, classical methods of testing hypotheses are used for this purpose. In this context, if the differences among residual variances or regression coefficients are "significant," the analysis of covariance model is no longer valid; if they are not "significant," we may proceed to do the analysis of real interest. However, even when the differences are not

"significant," it does not mean necessarily that the variables are homogeneous. It is hard to believe that either residual variances or regression coefficients are exactly the same among schools. The issue is, we think, whether or not this model fits sufficiently well to reality so that reasonably accurate conclusions may be reached. If the result of the testing of an hypothesis were regarded as a kind of index telling how different the groups are, it might be used for this purpose. But theoretically, the classical testing of hypothesis only serves to test whether the population values could reasonably be considered to be identical or not (or, more generally, whether some linear function of parameters is equal to some constant or not). It does not provide an index of "closeness of fit." We suggest that one way of solving the dilemma would be to first apply the m-group regression analysis to the data (Lindley and Smith, 1972; Novick, Jackson, Thayer and Cole, 1972). If the estimates obtained by this method, which are more reliable than the within group least squares estimates, are reasonably close to each other, we can go ahead and make a comparison of α_1's. If there are some outliers in terms of residual variances or regression coefficients, we seek another procedure for studying those schools and proceed with our analysis using the remaining schools.

Of course such a procedure for eliminating outliers does not solve all the problems. As a number of investigators have pointed out, whenever there are real differences in the pretest means, comparisons of groups using analysis of covariance leads to biased estimates of output measures. The bias results both from the unreliability of the input measures (Cronbach & Furby, 1970) and from the regression effects remaining after correction for unreliability (Campbell & Erlebacher, 1970).

To the extent that adjustments can be made for such biases, they may be applied in the context of the methods proposed here. To our knowledge, however, no attempt has been made to deal with the problem of small samples, and it is this problem that is the focus of this paper.

To minimize the problems associated with biases in the analysis of covariance, the data reported here are based on highly reliable measures, where the correlation between pretest and posttest measures is high, and where the differences in pretest means are small relative to the overlap in the several schools. Under such conditions, we think it is often meaningful to compare schools, under the assumption that all schools had students of equivalent initial abilities. Under such circumstances, we adopt the analysis of covariance model, i.e. Y_{ij}'s are independently distributed as a normal distribution with mean

$$\alpha + \sum_{k}^{p} X_{ijk}\,\beta_k \text{ and variance } \phi: \quad N(\alpha_i + \sum_{k}^{p} X_{ijk}\,\beta_k, \phi).$$

Prior Specification and Bayesian Solution

We will use a Bayesian method for estimating α's, and first we have to specify the prior density distributions. Before the investigation starts, we have virtually no knowledge about which school is really doing well or which school is doing rather poorly. Thus, we may not be able to distinguish a particular school i from the rest in terms of school effectiveness. Moreover, we may not be able to distinguish any pair of schools from any other pair, any three schools from any other three, and so on. In this case, the prior joint distribution for α's is invariant under any permutations of the subscripts. This is the property that DeFinetti (1964) has called exchangeability. Therefore,

α's are independently and identically distributed when the hyperparameters that determine the distribution of α's are given. Assuming normality,

$$(\alpha_i | \mu_\alpha, \phi_\alpha) \sim N(\mu_\alpha, \phi_\alpha) \ . \tag{5}$$

This exchangeable prior represents a situation before we start collecting any information. We do not know much about the β's and ϕ_α. This vague prior knowledge about β's is approximated by a uniform distribution over a range of interest, and we assume ϕ_α is distributed as an appropriate inverse chi-square distribution with some modest number of degrees of freedom, say 1. Further prior specifications are that the β's are distributed uniformly and ϕ is distributed as another appropriate inverse chi-square distribution. The same modal distribution and the same prior distributions were used as a prediction model and the Bayesian solution was discussed by Shigemasu (1976).

Approximately, the modal estimate of α_i is

$$\tilde{\alpha}_i \approx \tilde{\rho}_i \hat{\alpha}_i + (1 - \tilde{\rho}_i)\hat{\alpha}_. \ , \tag{6}$$

where

$$\tilde{\rho}_i = \frac{\tilde{\phi}_\alpha}{\tilde{\phi}_\alpha + n_i^{-1}\tilde{\phi}} \ ,$$

($\sim$ over the parameters denotes the Bayesian model estimate and n_i is the number of pupils in the i-th school,)

$\hat{\alpha}_i$ is a classical estimate, and

$$\hat{\alpha}_. = \frac{\Sigma n_i \hat{\alpha}_i}{\Sigma n_i} \ .$$

In this paper, we also take the modes of the joint posterior distribution as the estimates, but only the estimates of α's are of real interest. The criterion for assessing the various estimates is, of course, a loss function such as mean squared error, mean absolute error or zero-one loss. A loss function associated with modal estimates is zero-one loss (Novick and Jackson, 1974). Therefore if we are dealing with artificial data in which the "true" value is known, the various estimates can be evaluated in terms of a loss function. However, even without this evaluation, we believe that the loss associated with the Bayesian estimates is generally less than that using other estimates. The well known result by Stein (1956) implies that the usual types of least squares estimates in any linear model are not the best in that always other estimates, which are in the form of weighted averages that reduce the squared error loss can be found. The Bayesian approach with the exchangeability assumption is a natural way to obtain such estimates as minimize the squared error loss. We expect similarly that the zero one loss with the Bayesian modal estimates is smaller than that with the usual least squares method.

Comparison of Schools Within a System

The data for this analysis consist of test scores for two cohorts of pupils in 19 elementary schools in a midwestern city. The scores are for the five subtests--Vocabulary (V), Reading (R), Language (L), Work Study Skills (W), and Mathematics (M)--and for the Composite (C) of the Iowa Tests of Basic Skills administered in Grade 4 and again in Grade 6. Cohort I (n = 1002) took the tests first in January 1970 and

again in January 1972; Cohort II (n = 990) took the tests first in January 1971 and again in January 1973. The pupils for whom complete data were available represent approximately 90% of those tested in Grade 6. The cases were distributed among the schools as follows:

School	Cohort I	Cohort II
1	51	45
2	54	61
3	41	37
4	28	27
5	77	83
6	62	57
7	53	70
8	35	25
9	64	72
10	80	65
11	52	56
12	86	76
13	24	22
14	51	59
15	69	65
16	34	34
17	21	21
18	72	65
19	48	50

Selection of Covariates--Multiple Predictors or a Single Summative Predictor

The covariates available for this study are the five subtest and the composite scores of the ITBS administered in Grade 4. There are, of course, some other possible covariates such as latent traits of the pupils (e.g. I.Q. and motivation factor) and environmental variables (e.g. home and school conditions). But notice the inclusion of these measures as covariates makes it necessary to administer some other additional tests or to quantify the environmental factors, which may be impractical for the evaluator. However, it should be remembered that

such factors are already reflected in the initial achievement scores. Therefore, in this paper, we assume that the latent traits of the pupils and the environmental factors are adequately reflected in the difference of these pretest scores. In other words, we regard adjusting for differences in pretest scores to be equivalent to taking into account essentially all of the relevant preexisting differences.

Further, if most of the reliable variation of the five subtest scores is represented in the composite score, we may be able to use it as a single summative covariate rather than the whole set of five subtest scores. If it becomes routine to estimate a school effectiveness index, it will be much more convenient to identify and record at the time of testing only one score which was measured previously. A way of deciding whether or not a composite score can replace the five subtest scores may be to compare the residual variances with five subtest scores as predictors with those with only the composite score as predictor. Table 1 gives a comparison of residual variances in cross-validation data (one cohort) applying the estimates obtained in an analysis of the data for the other cohort. It is a moot point whether or not the residual variances reported in Table 1 differ sufficiently to warrent the use of the five covariates. Some may argue that the differences are considerable and we should use the five subtest scores. Others may argue that such a difference in the residual variance can be ignored. Our own judgment is that a composite score can be used as a covariate. However we reached this conclusion not only on the basis of the data in Table 1, but also by seeing the similarity between estimates of α's based on five subtests and on the composite score alone, as reported in the next section.

Results

Assuming that the analysis of covariance model is a reasonable one, we obtained both classical method least squares estimates and Bayesian modal estimates of α's. Tables 2, 3, 4, 5, 6 and 7 give these estimates when all x's = 45, the norm for the Grade 4 tests. Since we have two cohorts and two different sets of covariates, we have eight different estimates for each post test.

Tables

The most salient characteristics feature of the Bayesian results is that the estimates of α's for V and R are much more regressed toward the mean than W, M, and C. If we base our decision on the classical estimates, we might identify some schools, in which some teaching programs are supposedly very effective in building vocabulary, or developing reading ability. On the other hand, according to the Bayesian estimates, all schools are more or less the same in terms of V and R. An issue is whether or not the variation seen in the classical estimates are due only to the sampling error or whether they represent also real differences in achievement. We feel that the Bayesian estimates go better with our intuition. It is seldom observed that a public elementary school will employ very effective different educational program in developing the vocabulary of pupils. Rather, vocabulary responses seem to reflect primarily the developmental level attributable to the general stimulation of the total culture. Furthermore, for these particular schools, there had been a common emphasis on the reading program, fully supported by the central administration of the system.

Classical Vs. Bayesian Estimates

In this paper, we have estimates for two cohorts separated by only one year. Here let us consider the characteristics of the two estimation methods by comparing the estimates of one cohort with those of another. Table 8 gives the mean squares of the differences between the estimates of the α's for the two cohorts. Of course, one might expect some change between the two cohorts; for example, there may be changes in the teaching staff. However, it is very likely that the school and home conditions in general will remain more or less the same between the two cohorts. From Table 8, we see that the Bayesian estimates are less different from cohort to cohort than the classical estimates. If we conclude after careful investigation that not much change should be expected between two cohorts, this fact would suggest that the Bayesian estimates are the more accurate. However, we should be very careful about this point. The more regressed toward the central tendency, the smaller the mean square of differences will be, and so this fact may only tell us that the Bayesian estimates are regressed. Let us examine Table 8 more closely. According to the classical estimates, the ranks of the six mean square of differences are M-L-V-R-W-C (with five subtest scores as covariates) and M-V-L-R-W-C (with the composite score as covariate), whereas for the Bayesian estimates, the ranks are L-M-W-C-V-R (5 subtests) and L-M-W-C-V-R (composite). Our feeling is that the subtests which are most resistant to change are Vocabulary, Reading and Workstudy: Vocabulary, because the pupils build vocabulary from almost any source; Reading, because the reading programs are receiving primary emphasis in all schools; and Workstudy Skills because no schools have specific courses that focus on these skills. The composite score, which

combines the characteristics of the five subtests, may well show an average amount of variation from cohort to cohort. The Bayesian estimates seem to be more consistent with these intuitions.

On the other hand, the correlations between the α values for the two cohorts are almost the same whether the α values are based on classical or on Bayesian procedures (Table 9). This shows that when we are interested only in ranking schools, the Bayesian estimation does not improve the situation, because the shrinkage rate is about the same across schools. (When we employ the more general model, in which heterogeneous residual variances are allowed, the rate of shrinkage is a function of the residual variance of each school, and we may expect some improvement in terms of correlation, too.)

In order to see the relative position of each school more easily, $\alpha_i^* = \alpha_i - \alpha_.$ $(\alpha_. = m^{-1}2\alpha_.)$ may be more appropriate than α_i. Suppose that $\alpha_o = 1.25$ is judged to be a critical value, i.e. if $\alpha_i^* > \alpha_o$, school i is judged to be doing very well for some reason, and if $\alpha_i^* < -\alpha_o$, school i may be suspected of having some deficiency in teaching. An intuitive way of judging whether schools are doing well or poorly is to see whether or not the modal estimate: $\tilde{\alpha}_i^* = \tilde{\alpha}_i - \tilde{\alpha}_.$ is greater than α_o or smaller than $-\alpha_o$. Table 10 gives the number of schools in which we had different decisions for two cohorts. Table 11 gives the number of schools in which our judgment is $\alpha_i > \alpha_o$ for one cohort and $\alpha_i < -\alpha_o$ for another cohort. The differences between classical estimates and Bayesian estimates were generally small except for V and R. But the difference in the case of V may be a compelling reason for using the Bayesian instead of the classical estimates.

The Analysis of Covariance model that is used in this paper for comparing the school effectiveness suffers from three limitation.

1. All that was done in this paper was to derive a reliable index of school effectiveness, and for this purpose we used the modal estimate. However, the modal estimate is only a measure of central tendency of a posterior distribution. If we wish to make decisions on the basis of the data we need to know the marginal distribution of α's. For example, the researcher may be required to decide: (i) whether or not the effectiveness of school i can be said to be greater than some constant, or (i) whether or not the effectiveness of school i can be said to be greater than that of school ℓ ($i \neq \ell$), or (iii) which teaching method is recommended if the cost is taken into account. All these decisions involve the marginal posterior distribution and a loss function or utility. An investigation of this distribution is under development.

2. So far we have discussed the problem assuming that the x's are given, and we confined our attention to the distribution of y_{ij} conditional on x_{ij}, and not that of the x's themselves. However, we have every reason to believe that the x's are fallible, and it might be better to compare school effectiveness, taking into account the differences among true values (τ_{ij}) rather than among the x's themselves. More specifically we are interested in the following situation:

$$x_{ijk} \sim N(\tau_{ijk}, \phi_{xk}) \quad (\text{for } i = 1, \ldots, m,\ j = 1, \ldots, n_i,\ k = 1, \ldots, p) \tag{7}$$

$$\tau_{ijk} \sim N(E_{ik}, \phi_{\tau k}) \quad (\text{for } i = 1, \ldots, m,\ j = 1, \ldots, n_i,\ k = 1, \ldots, p) \tag{8}$$

$$Y_{ij} | \tau_{ij1}, \ldots \tau_{ijp} \sim N(\alpha_i + \beta_1 \tau_{ij1} + \ldots + \beta_p \tau_{ijp}, \phi) \ . \tag{9}$$

We feel it is certainly worthwhile to obtain the estimates of α_i under these assumptions, and to compare them with the estimates given in this paper.

They may or may not differ to any practical degree.

3. The third point, which involves the homoscedascity assumption, was discussed briefly earlier in the paper. In some situations, it may be desirable to apply the analysis of covariance model assuming the residual variances are different. In this case, ϕ's will be assumed to be exchangeable, and, as a result, the collateral information can be made use of.

Our eventual goal is to report a school effectiveness index to each school in order to stimulate study of teaching practices. We feel meanwhile reports such as these should be viewed as suggestive only, not as definitive. Certainly, the three possible improvements should be examined carefully before implementing the reporting system.

REFERENCES

Campbell, D. T. & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education tend to underestimate effects. In C. A. Bennett & A. Lumsdaine (Eds.) Evaluation and experiment: some critical issues in assessing social programs. New York: Academic Press, 1975.

Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Cronbach, L. J. et al. Analysis of covariance: angel of salvation or temptress and deluder? Stanford, CA: Occasional Papers of the Stanford Evaluation Consortium, Department of Education, Stanford University, February 1976.

Cronbach, L. J. & Furby, L. How should we measure "change" -- or should we? Psychological Bulletin, 1970, 74, 68-80.

De Finetti, B. Foresight: its logical laws, its subjective sources. In H. E. Kegburg, Jr. & H. E. Smokler (Eds.) Studies in subjective probability. New York: Wiley, 1964.

Dyer, H. S. Toward objective criteria of professional accountability in the schools of New York City. Phi Delta Kappan, 1970, 52, 206-211.

Dyer, H. S., Linn, R. L. & Patton, M. J. Feasibility study of educational performance indicators: final report to New York State Education Department. Princeton, N. J.: Educational Testing Service, 1967. (ERIC Document Reproduction Service No. ED 079353.)

Dyer, H. S., Linn, R. L. & Patton, M. J. A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. American Educational Research Journal, 1969, 6, 591-605.

Elashoff, J. D. Analysis of Covariance: a delicate instrument. American Educational Research Journal, 1969, 6, 383-401.

Forsyth, R. Some empirical results related to the stability of performance indicators in Dyer's student change model of an educational system. Journal of Educational Measurement, 1973, 10(1), 7-12.

Horst, D. P., Tallmage, G. K., & Wood, C. T. A practical guide to measuring project impact on student achievement. Washington, DC: U.S. Dept. of Health, Education, and Welfare, 1975.

Lindley, D. V. and Smith, A. F. M. Bayesian estimates for the linear model. Journal of Royal Statistical Society, Series B, 1972, 34, 1-41.

Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin Company, 1953.

Lord, F. M. A paradox in the interpretation of group comparisons. Psychological Bulletin, 1967, 68, 304-305.

Novick, M. R., Jackson, P. H., Thayer, D. T. & Cole, N. S. Application of Bayesian methods to the prediction of educational performance. The British Journal of Mathematical Psychology, 1972, 25, 33-50.

Novick, M. R. and Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw Hill, 1974.

Porter, A. C. The effects of using fallible variables in the analysis of covariance. (Doctoral dissertation, University of Wisconsin). Dissertation Abstracts International, 1967, 28, 3517B. (University Microfilms No. 67-12147).

Shigemasu, K. Development and validation of a simplified m-group regression model. Journal of Educational Statistics, 1976, 1 (2), 157-180.

Stein, C. M. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Proc. Third Berkeley Symposium. Berkeley: University of California Press, 1956.

Table 1

A Comparison of Residual Variances for Cross Validation Data
(Cohort I is Cross Validation Sample 1; Cohort II, Sample 2)

Test	Sample	Least Square Estimates		Bayesian Estimates	
		Five Subtests	Composite	Five Subtests	Composite
V	1	90.05	103.50	86.75	100.06
	2	90.69	104.28	88.55	101.90
R	1	71.68	77.12	70.94	75.83
	2	73.59	78.56	73.64	78.41
L	1	61.34	74.50	60.76	73.59
	2	65.49	81.11	65.18	80.68
W	1	61.44	65.15	60.68	64.20
	2	60.03	69.06	59.41	68.56
M	1	56.95	62.13	56.69	61.65
	2	56.27	67.96	55.35	66.94
C	1	30.39	30.29	30.05	30.24
	2	32.12	32.11	31.91	31.92

Table 2

Bayesian and Classical Estimates of ITBS Vocabulary Grade 6 Scores
for Two Cohorts of Pupils in 19 Schools,
Assuming Grade 4 Scores of 45

	1971-1972				1972-1973			
	BAYESIAN ESTIMATES		CLASSICAL ESTIMATES		BAYESIAN ESTIMATES		CLASSICAL ESTIMATES	
	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE
1	65.09	65.18	65.74	65.72	65.93	65.95	66.23	65.56
2	64.72	64.94	62.48	63.93	65.76	65.93	65.19	65.55
3	65.10	65.13	65.98	65.45	65.94	66.11	66.36	66.91
4	65.18	65.36	67.68	68.57	66.05	66.09	67.91	66.96
5	65.11	65.22	65.63	65.75	65.94	66.07	66.15	66.32
6	64.82	64.78	63.67	63.02	66.05	66.36	66.86	68.27
7	65.07	65.08	65.60	64.95	65.63	65.86	64.58	65.21
8	65.30	65.37	68.70	67.98	65.72	65.83	63.76	63.62
9	65.04	65.09	65.34	65.09	65.91	65.89	66.02	55.42
10	65.09	65.05	65.53	64.83	65.25	65.38	62.37	62.46
11	65.02	65.16	65.14	65.61	65.52	65.73	63.65	64.18
12	64.69	64.88	63.17	64.03	65.72	65.76	65.10	64.76
13	65.10	65.15	66.69	65.89	65.91	66.05	66.26	66.55
14	64.74	64.73	62.68	62.29	65.73	65.93	65.00	65.53
15	64.79	64.81	63.52	63.39	66.53	66.57	69.46	69.19
16	65.13	65.42	66.65	68.62	66.02	66.21	67.19	68.00
17	65.12	65.23	67.18	67.49	65.79	65.82	64.43	63.14
18	64.83	64.72	63.88	62.97	65.92	66.02	66.06	66.06
19	65.26	65.49	67.46	68.24	66.47	66.65	69.92	70.49

Table 3

Bayesian and Classical Estimates of ITBS Reading Grade 6 Scores
for Two Cohorts of Pupils in 19 Schools,
Assuming Grade 4 Scores of 45

	1971-1972				1972-1973			
	BAYESIAN ESTIMATES		CLASSICAL ESTIMATES		BAYESIAN ESTIMATES		CLASSICAL ESTIMATES	
	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE
1	66.41	66.57	67.11	67.30	65.83	65.80	66.37	65.82
2	65.83	65.72	63.86	63.92	65.39	65.31	64.68	64.63
3	66.11	65.98	65.17	64.60	65.43	65.23	64.29	64.09
4	66.63	66.89	69.72	69.98	65.74	65.55	65.99	64.96
5	66.64	66.79	67.85	67.74	66.05	66.27	66.75	66.60
6	66.36	66.44	66.80	66.73	65.91	66.27	66.56	66.89
7	66.15	66.09	65.69	65.42	66.19	66.72	67.38	67.83
8	66.35	66.45	67.03	67.00	65.56	65.24	64.73	63.71
9	66.27	66.42	66.30	66.61	65.51	65.18	65.12	64.46
10	66.00	65.87	65.28	65.02	64.63	63.98	62.15	61.83
11	65.89	65.80	64.10	64.15	65.36	65.18	64.41	64.26
12	66.04	66.18	65.45	65.90	65.56	65.39	65.26	64.87
13	66.55	66.69	69.35	68.99	65.93	66.07	67.61	67.24
14	65.93	65.85	64.41	64.36	65.69	65.91	65.77	66.02
15	66.33	66.27	66.60	66.17	66.06	65.99	66.91	66.12
16	66.70	67.15	69.78	70.92	66.35	67.03	69.37	69.75
17	66.39	66.48	67.70	67.49	65.25	64.63	61.91	60.64
18	66.08	65.95	65.51	65.14	65.32	65.04	64.40	64.10
19	66.51	66.93	67.80	68.87	66.43	67.26	68.86	69.42

Table 4

Bayesian and Classical Estimates of ITBS Language Grade 6 Scores
for Two Cohorts of Pupils in 19 Schools,
Assuming Grade 4 Scores of 45

	1971-1972				1972-1973			
	BAYESIAN ESTIMATES		CLASSICAL ESTIMATES		BAYESIAN ESTIMATES		CLASSICAL ESTIMATES	
	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE
1	68.66	68.49	69.05	69.01	66.04	67.81	66.32	68.23
2	62.33	62.05	62.00	61.49	62.68	63.95	62.65	63.81
3	62.11	64.01	61.63	63.69	61.65	62.32	61.40	61.78
4	63.97	64.04	63.70	63.61	63.46	65.54	63.41	65.63
5	68.44	67.38	68.69	67.61	70.09	70.82	70.51	71.30
6	63.42	63.20	63.26	62.91	62.88	61.96	62.85	61.59
7	61.95	63.06	61.59	62.70	65.84	66.22	66.05	66.23
8	65.88	66.05	65.99	66.22	61.89	63.62	61.53	63.19
9	62.71	62.92	62.47	62.57	62.70	64.29	62.68	64.21
10	66.04	66.34	66.11	66.47	60.99	62.27	60.84	61.97
11	62.58	64.29	62.26	64.08	60.58	61.89	60.34	61.47
12	63.94	62.73	63.85	62.46	59.71	60.71	59.50	60.30
13	67.89	68.03	68.50	68.94	67.40	68.13	68.19	69.08
14	64.65	64.29	64.58	64.10	61.97	63.27	61.88	63.06
15	67.54	68.42	67.73	68.81	63.00	64.67	62.98	64.62
16	72.67	72.35	73.93	74.10	72.55	72.58	73.81	74.08
17	61.72	62.75	60.71	61.60	60.80	63.04	60.11	62.33
18	68.81	69.15	69.09	69.60	65.76	66.30	65.95	66.42
19	66.73	65.90	66.89	65.97	67.92	69.37	68.39	69.96

Table 5

Bayesian and Classical Estimates of ITBS Work-Study Skills Grade 6 Scores for Two Cohorts of Pupils in 19 Schools, Assuming Grade 4 Scores of 45

	1971-1972				1972-1973			
	BAYESIAN ESTIMATES		CLASSICAL ESTIMATES		BAYESIAN ESTIMATES		CLASSICAL ESTIMATES	
	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE
1	66.12	65.99	66.09	65.96	66.11	64.98	66.12	64.92
2	62.40	62.05	61.26	60.81	64.25	63.13	63.89	62.67
3	63.67	63.44	62.67	62.36	65.50	64.81	65.29	64.69
4	66.51	66.24	66.65	66.31	65.85	65.20	65.73	65.19
5	67.78	67.83	68.10	68.22	70.67	69.53	71.34	70.23
6	65.99	66.33	65.98	66.42	66.70	66.12	66.83	66.35
7	64.95	64.94	64.60	64.61	65.38	64.46	65.27	64.17
8	68.00	68.02	68.80	68.92	67.75	66.76	68.56	67.58
9	68.66	68.77	69.27	69.43	69.12	69.10	69.64	69.82
10	65.35	65.46	65.21	65.37	61.05	60.04	60.11	58.99
11	64.32	63.70	63.73	62.92	63.97	62.97	63.50	62.43
12	66.48	66.47	66.54	66.54	64.89	64.22	64.69	64.05
13	67.57	67.61	68.42	68.60	66.05	65.38	66.04	65.49
14	64.47	64.72	63.96	64.29	65.24	63.56	65.06	63.19
15	67.53	67.35	67.82	67.66	67.95	67.38	68.30	67.82
16	69.69	68.84	71.25	70.12	67.84	67.07	68.47	67.79
17	65.04	64.98	64.12	64.05	65.31	64.80	64.86	64.53
18	66.40	66.67	66.45	66.81	64.68	64.26	64.41	64.07
19	68.46	68.37	69.22	69.11	67.61	65.72	67.98	65.86

Table 6

Bayesian and Classical Estimates of ITBS Mathematics Skills Grade 6 Scores for Two Cohorts of Pupils in 19 Schools, Assuming Grade 4 Scores of 45

	1971-1972				1972-1973			
	BAYESIAN ESTIMATES		CLASSICAL ESTIMATES		BAYESIAN ESTIMATES		CLASSICAL ESTIMATES	
	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE
1	59.48	59.52	59.03	58.96	63.60	62.62	63.45	62.38
2	61.34	61.28	61.09	60.97	62.30	60.64	62.00	60.06
3	63.52	63.12	63.47	63.03	61.97	61.51	61.38	60.86
4	65.43	65.02	65.71	65.35	64.88	64.38	65.13	64.85
5	65.10	65.63	65.17	65.81	66.29	65.22	66.52	65.49
6	65.30	65.32	65.41	65.51	65.27	65.41	65.39	65.83
7	59.10	59.26	58.62	58.70	62.70	61.60	62.46	61.11
8	60.25	60.46	59.70	59.83	63.17	62.71	62.76	62.35
9	60.90	60.49	60.66	60.15	63.38	63.13	63.27	63.07
10	61.56	61.87	61.40	61.72	60.58	59.92	60.02	59.22
11	63.61	62.39	63.59	62.22	64.92	63.51	65.07	63.56
12	65.22	65.21	65.30	65.34	64.31	63.97	64.32	64.06
13	66.59	66.40	67.16	67.18	61.25	61.17	59.92	59.84
14	62.29	62.52	62.12	62.37	63.89	62.22	63.83	61.95
15	70.20	70.04	70.67	70.68	70.41	69.75	71.33	71.00
16	68.72	67.26	69.46	67.98	65.20	63.84	65.49	64.01
17	63.23	62.95	63.07	62.71	66.15	65.56	67.04	66.89
18	64.17	64.33	64.19	64.39	62.80	62.94	62.58	62.84
19	68.27	66.92	68.76	67.38	66.90	64.07	67.41	64.23

Table 7

Bayesian and Classical Estimates of ITBS Composite Grade 6 Scores
for Two Cohorts of Pupils in 19 Schools,
Assuming Grade 4 Scores of 45

	1971-1972				1972-1973			
	BAYESIAN ESTIMATES		CLASSICAL ESTIMATES		BAYESIAN ESTIMATES		CLASSICAL ESTIMATES	
	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE
1	65.42	65.41	65.42	65.41	65.63	65.34	65.73	65.42
2	62.76	62.84	62.17	62.26	63.92	63.63	63.70	63.37
3	64.17	64.21	63.79	63.84	64.10	64.01	63.77	63.70
4	66.31	66.35	66.66	66.73	65.59	65.46	65.74	65.63
5	66.87	66.80	67.08	67.02	67.89	67.63	68.23	67.98
6	65.05	64.96	64.99	64.88	65.57	65.61	65.62	65.73
7	63.67	63.72	63.29	63.35	65.20	65.03	65.21	64.96
8	65.91	65.87	66.07	66.02	64.50	64.32	64.17	64.01
9	64.90	64.87	64.80	64.76	65.32	65.35	65.34	65.40
10	64.76	64.73	64.67	64.65	61.63	61.45	61.03	60.83
11	64.09	64.12	63.77	63.81	63.62	63.43	63.31	63.11
12	64.97	64.96	64.91	64.89	63.99	63.83	63.80	63.65
13	67.25	67.17	68.11	68.01	65.46	65.42	65.57	65.62
14	63.91	63.85	63.56	63.49	64.43	64.12	64.29	63.95
15	67.05	67.10	67.32	67.39	67.41	67.40	67.76	67.73
16	69.02	69.11	70.24	70.37	68.05	67.90	68.96	68.83
17	64.98	65.05	64.70	64.81	64.18	64.03	63.64	63.49
18	65.79	65.75	65.84	65.80	64.72	64.71	64.62	64.66
19	67.48	67.39	67.98	67.86	67.90	67.45	68.50	67.98

Table 8

Mean Squares of Differences between Estimates of Grade 6 ITBS Scores for Two Cohorts of Pupils in 19 Schools

	BAYESIAN ESTIMATES		CLASSICAL ESTIMATES	
	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE
V	0.88	0.96	6.29	7.43
R	0.44	0.86	4.09	6.18
I.	6.76	5.37	7.59	6.85
W	2.52	3.54	4.06	5.23
M	5.73	4.92	8.14	7.78
C	1.33	1.38	2.05	2.06

Table 9

Correlations between Estimates of ITBS Grade 6 Scores for Two Cohorts of Pupils in 19 Schools

	BAYESIAN ESTIMATES		CLASSICAL ESTIMATES	
	5 SUBTESTS	COMPOSITE	5 SUBTESTS	COMPOSITE
V	0.03	0.08	0.05	0.06
R	0.62	0.57	0.51	0.47
I	0.75	0.71	0.77	0.73
W	0.68	0.68	0.68	0.68
M	0.63	0.64	0.58	0.58
C	0.75	0.76	0.75	0.76

Table 10

Frequency of Inconsistency of Classifications of 19 Schools when Classifications are Based on Data from Two Different Cohorts. (Categories: $\alpha_i^* > 1.25$; $1.25 > \alpha_i^* > -1.25$; $\alpha_i^* < -1.25$)

	Bayesian Estimates		Classical Estimates	
	5 Subtests	Composite	5 Subtests	Composite
V	0	0	13	10
R	0	3	7	8
L	7	9	6	8
W	6	6	8	6
M	13	11	11	10
C	5	6	6	6

Table 11

Frequency with which 19 Schools are Classified at One Extreme on the Basis of Data from One Cohort and at the Other Extreme on the Basis of Data from a Second Cohort. (Categories: $\alpha_i^* > 1.25$; $1.25 > \alpha_i^* > -1.25$; $\alpha_i^* < -1.25$)

	Bayesian Estimates		Classical Estimates	
	5 Subtests	Composite	5 Subtests	Composite
V	0	0	3	4
R	0	0	0	0
L	1	0	1	0
W	0	0	0	0
M	1	1	1	1
C	0	0	0	0